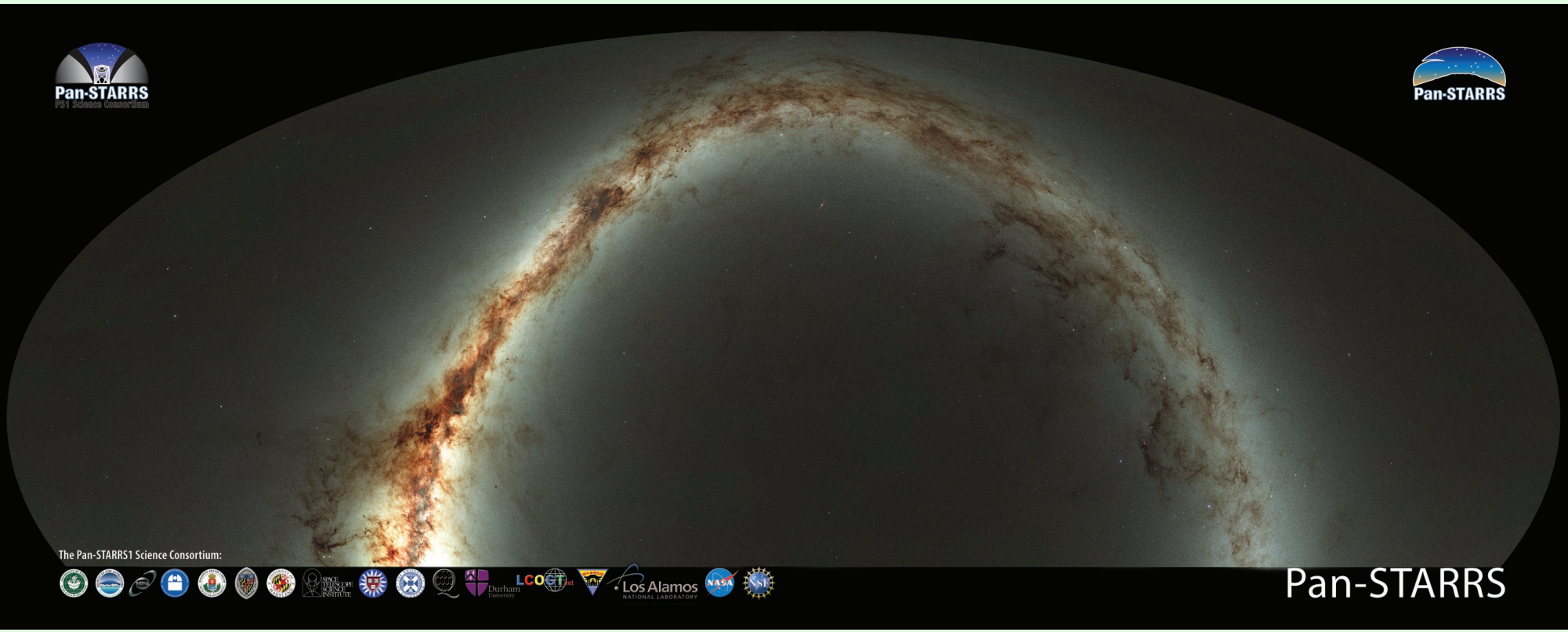
Lessons from a Large Survey: The First Decade of Pan-STARRS Observations

Christopher Waters

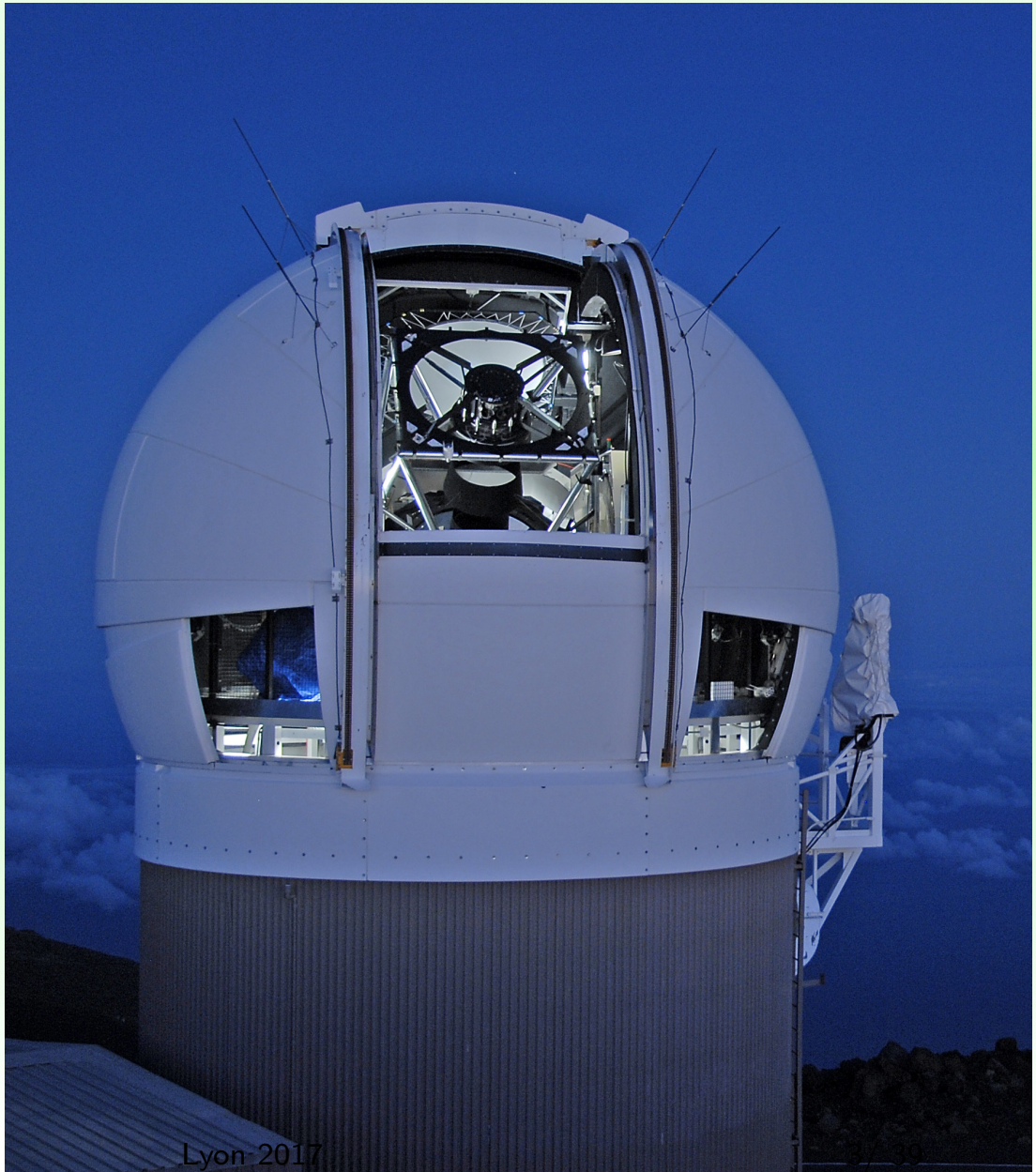
Institute for Astronomy

Introduction



Pan-STARRS1

- 1.8m telescope at the summit of Haleakala on Maui
- 3.2 degree field of view
- Equipped with GPC1
  - 1.4 gigapixel camera
  - $0\farcs26$ / pixel
- First light August 2007, full pre-survey science operations December 2009
- Science Consortium surveys operated from May 2010 through March 2014

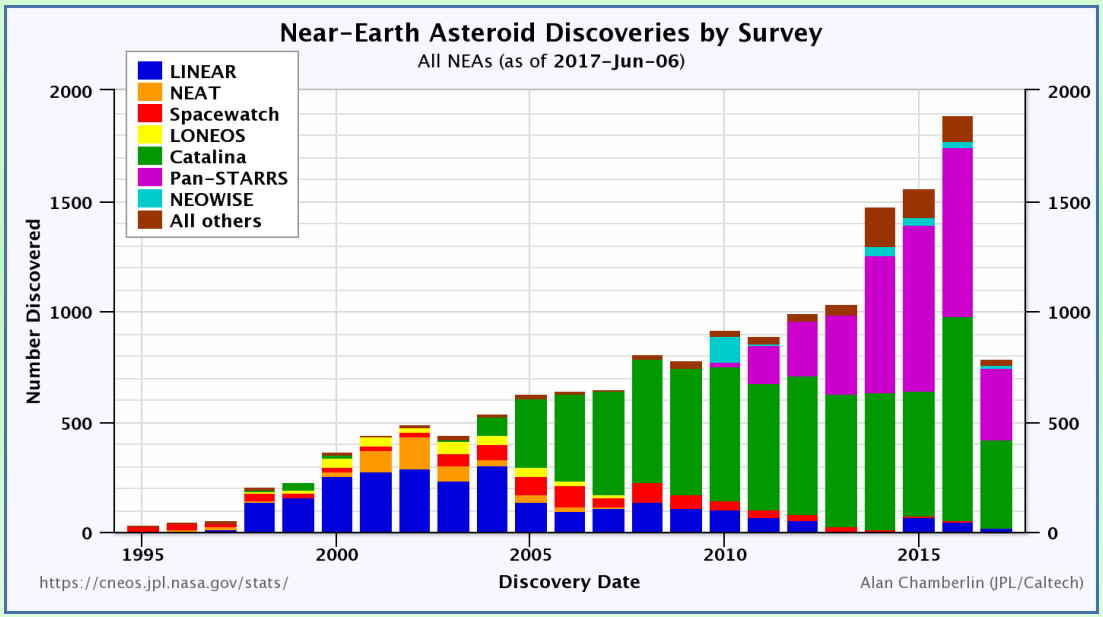


Science Consortium Surveys:

- 3π: full sky north of $\delta > -30$; *grizy*
- Medium Deep: 10 single pointing fields, 8-point dither; *grizy*
- Moving Objects: focused on ecliptic; *w*
- Stellar Transit: high time resolution; *i*
- M31: nightly monitoring; mainly *ri*

Continuing NEO Mission:

- Since 2014, Pan-STARRS has been continuing in search for NEO
- Responsible for roughly half of discoveries
- Should increase when PS2 and GPC2 fully come online



Data Release

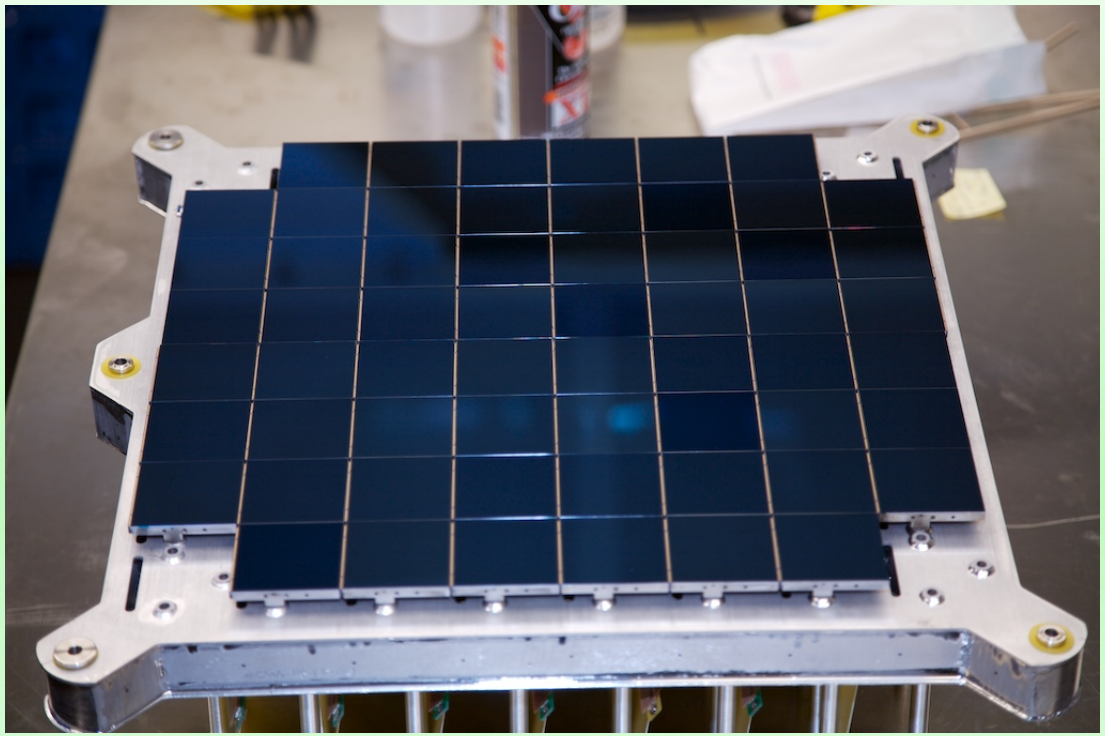
- MAST Archive at STScI has all 3π stack and warp images.
- Currently, only average object parameters and stack images and photometry available
- Full final catalog and individual frames will be released in DR2
  - 3e9 objects
  - 8.5e10 single epoch detections
  - 3.8e11 forced warp detections
- Medium Deep data coming in the future as well

What have we learned?

- Many good design decisions were made early
- We still would not have been able to do this data release without near constant development
- At the start of science observations, processing had to be manually launched

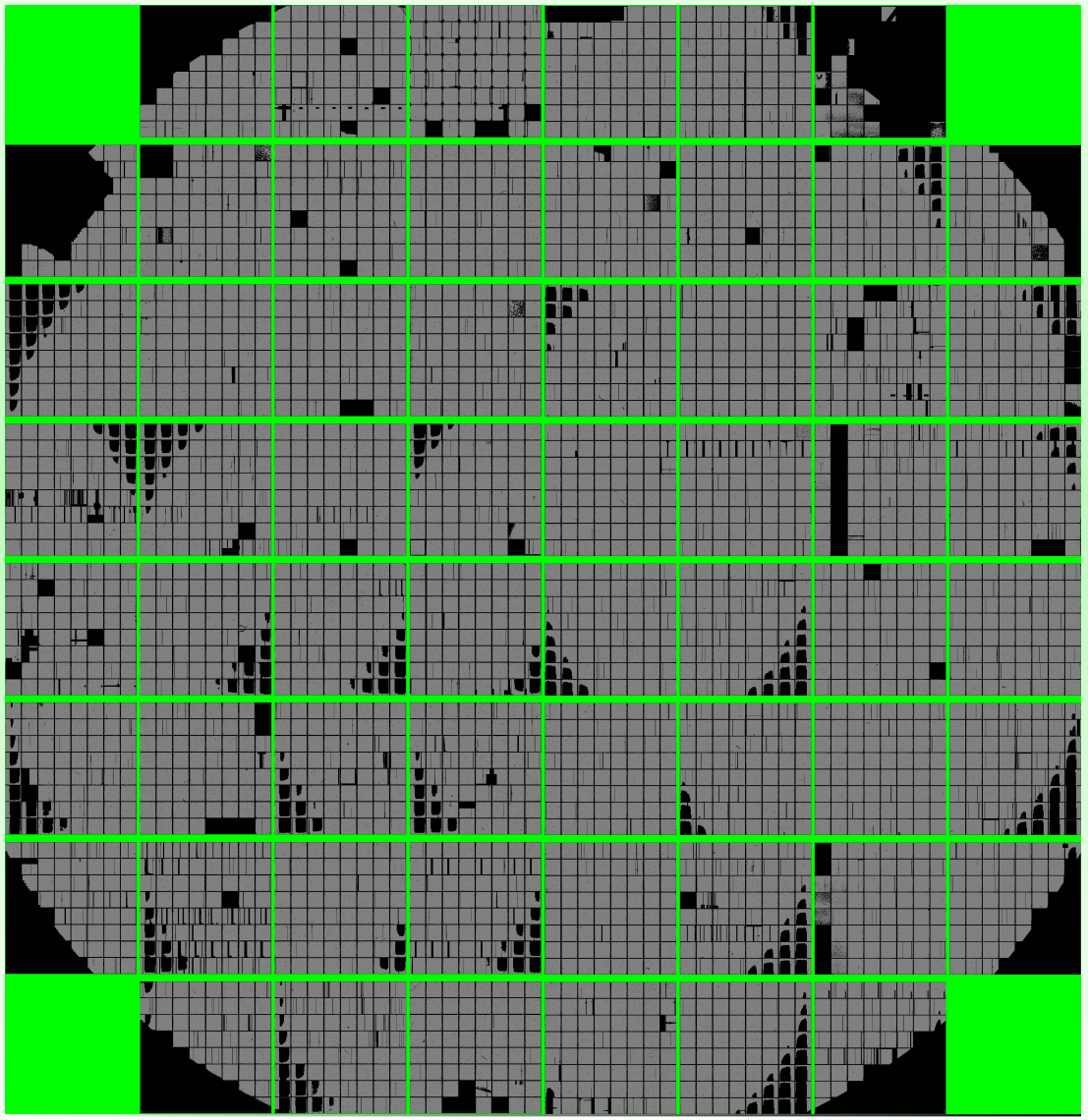
Lesson 1: Parallelization

- GPC1 is designed to be parallel
  - 60 OTAs
  - Each contain 64 readout cells
  - 10.3s overhead
- Downside is that the usable area of the focal plane is decreased



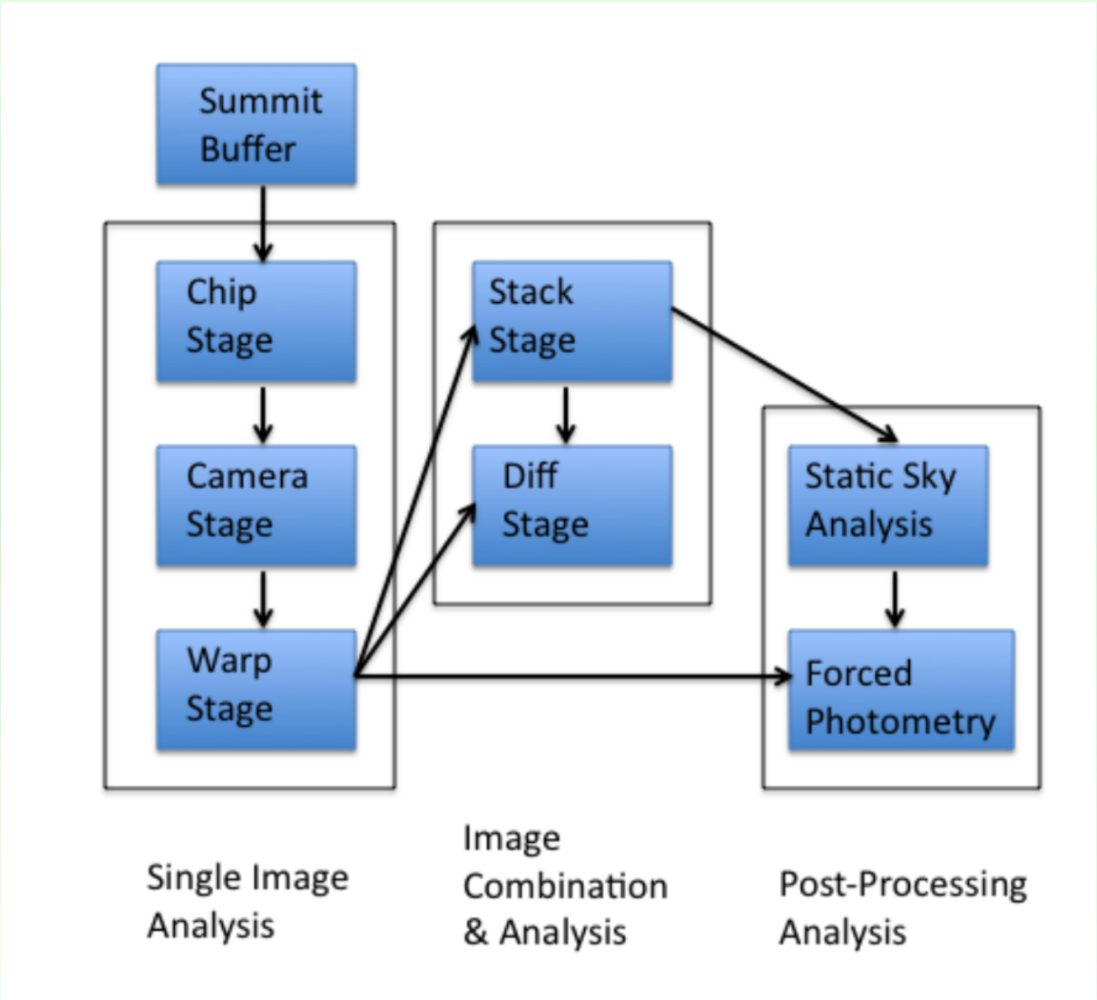
Lesson 1: Parallelization

- GPC1 is designed to be parallel
  - 60 OTAs
  - Each contain 64 readout cells
  - 10.3s overhead
- Downside is that the usable area of the focal plane is decreased
- Average fill factor around 80%



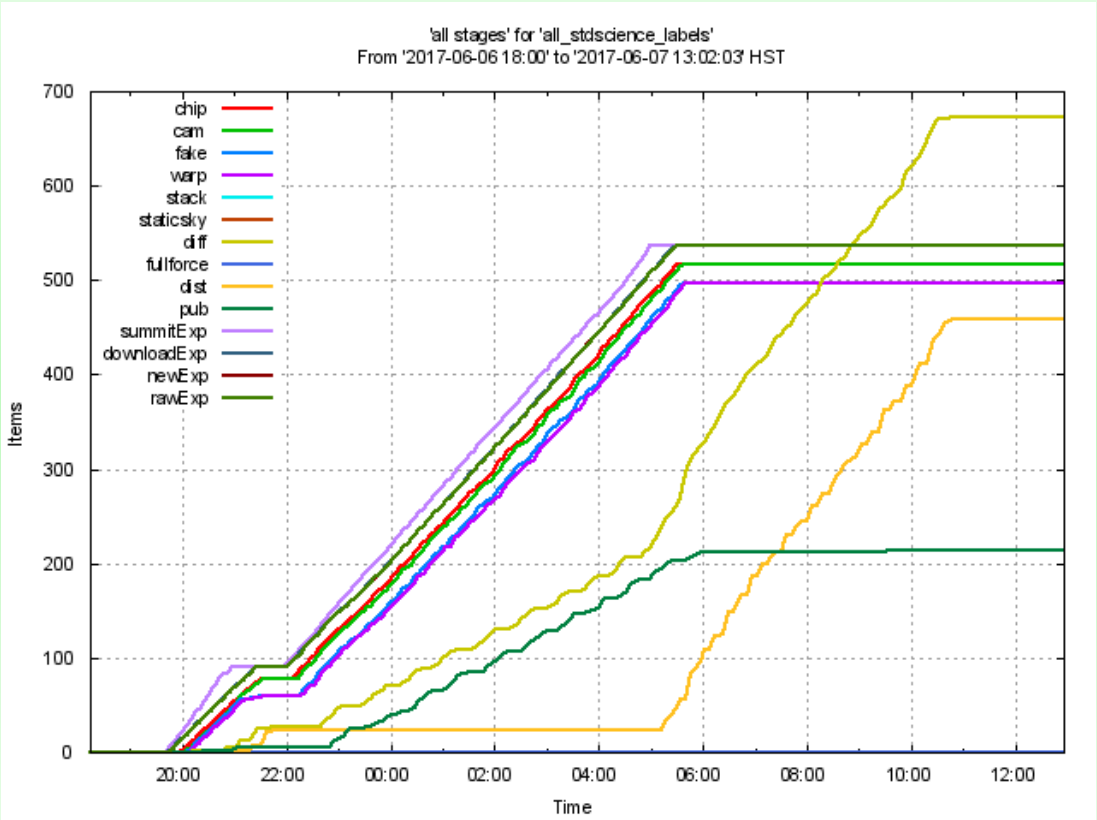
Processing

- Exposures are split across processing cluster
- Each OTA reduced separately
- Warp processing is performed on 6000×6000 pixel skycell images
- Stacks and differences work the same.



Current Pipeline

- Automatically downloads exposures from the summit to data center
- Registers new exposures in processing database
- Processing for new data occurs in real time, completes by morning
- Allows moving object pipeline to identify potential sources for followup



Parallelize Storage

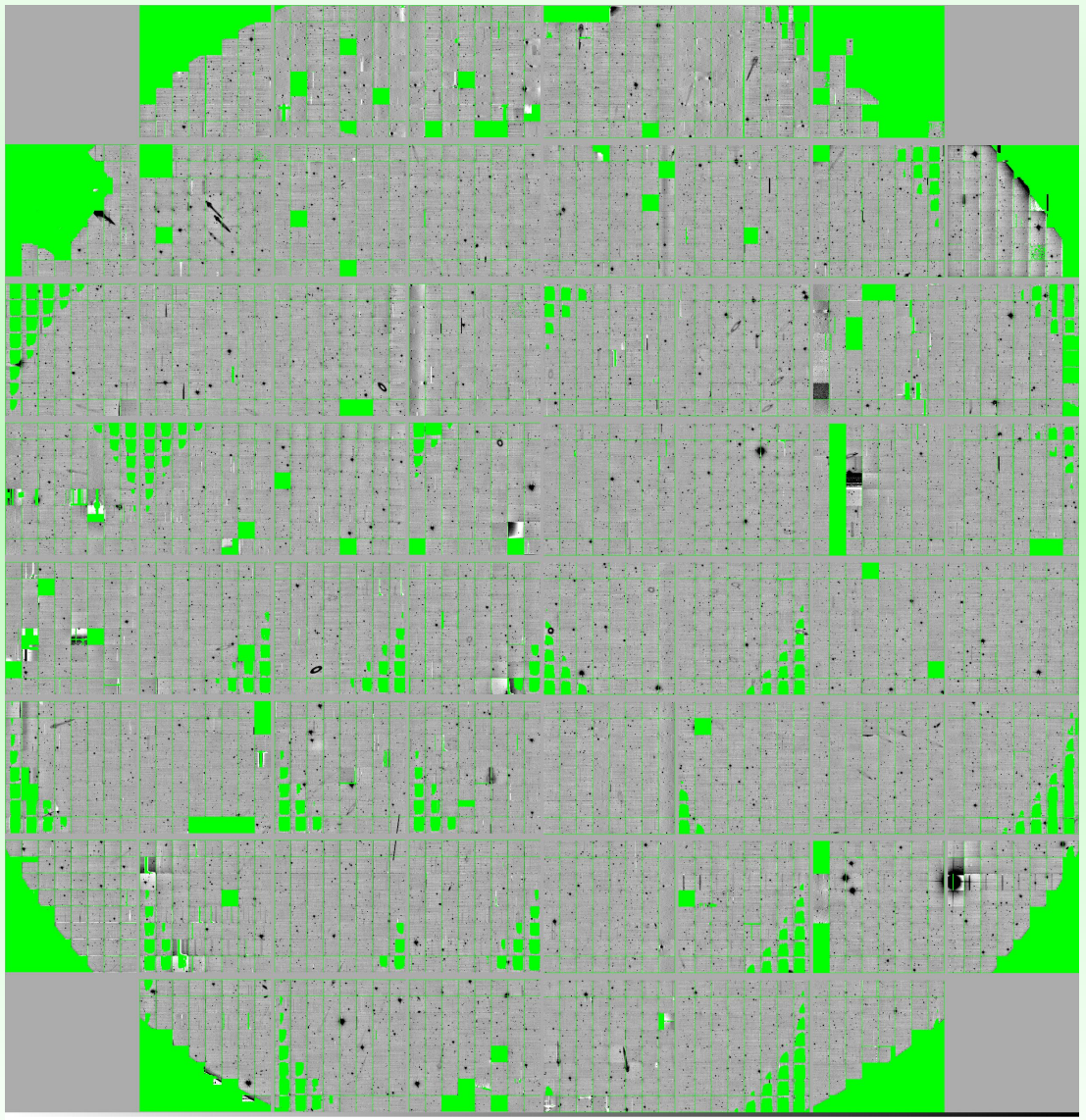
- Use “Nebulous” system for file storage
- Storage nodes mountable via NFS
- Database stores mapping between “neb name” and disk location
- Queries are fast, and file retrieval largely network limited
- Originally focused on targeting data, but 10G networks loosen this constraint
- Multiple copies of important data to allow for host unreliability
- Limited location awareness

Object and Detection Database

- Internally use Desktop Virtual Observatory (DVO) catalog
- Not parallel at the start of observing, but clearly essential
- Does final astrometric and photometric calibrations across the entire catalog of detections
- Current 70 TB catalog split into 100 subsets on 13 machines
- Full database calibration now requires 9.8 TB of memory and a day of processing on 100 hosts

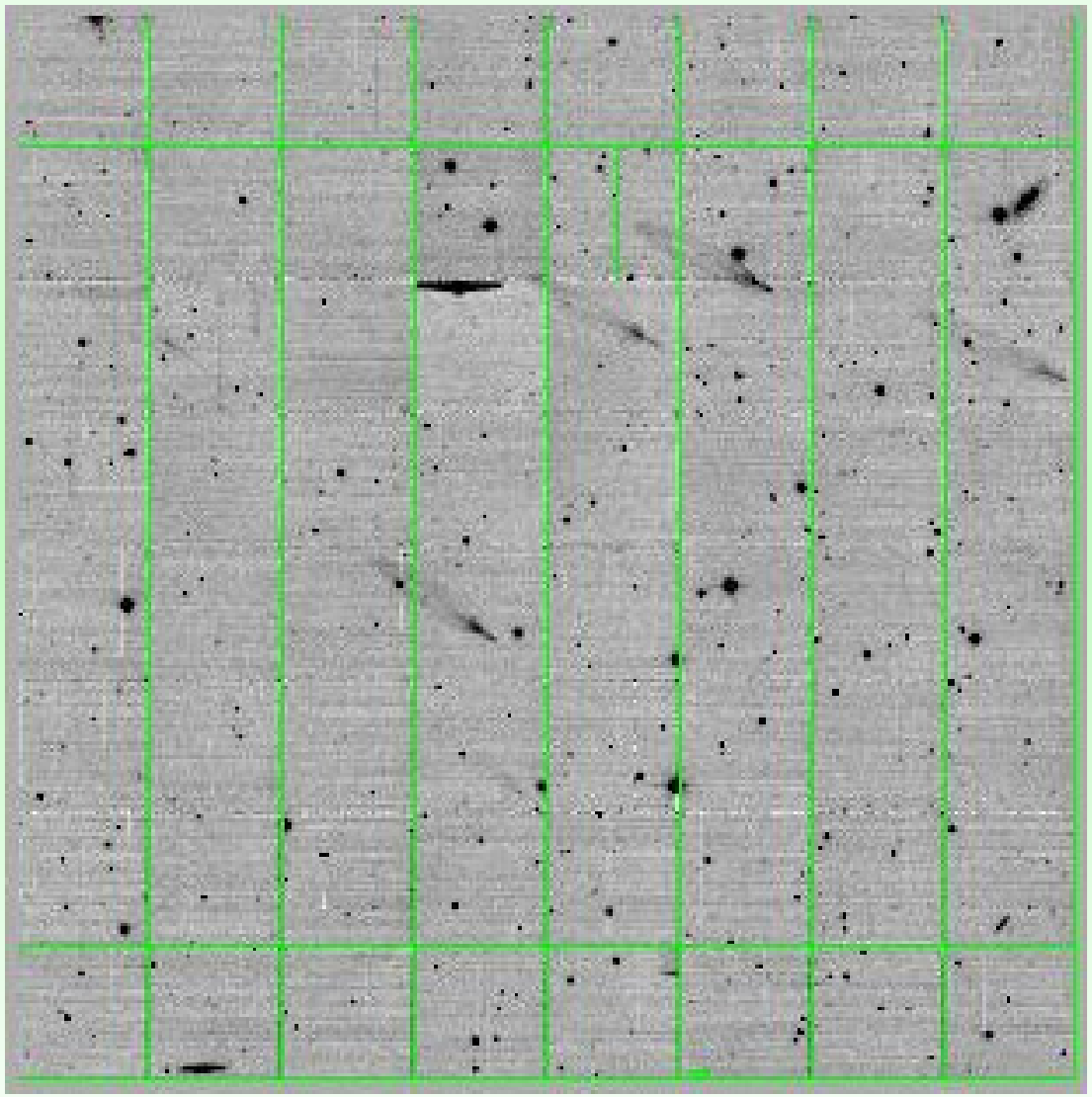
Make the code generic

- IPP code makes a minimum assumptions about the data as possible
- A “camera” consists of an FPA that contains one or more chips, each of which contains one or more cells, each of which contains one or more readout
- This accounts for GPC1, and allows video guide cells with multiple readouts



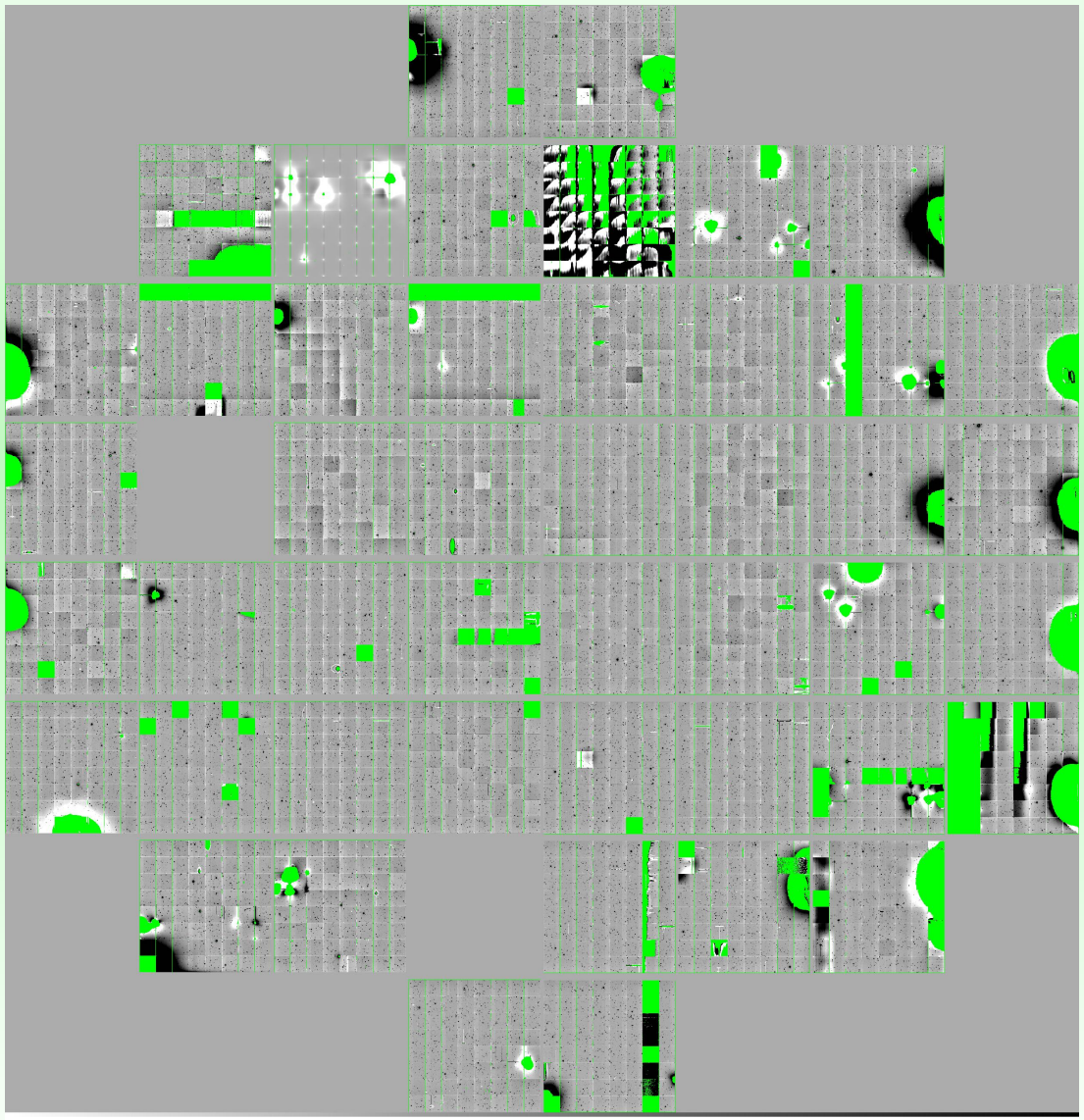
Make the code generic

- IPP code makes a minimum assumptions about the data as possible
- A “camera” consists of an FPA that contains one or more chips, each of which contains one or more cells, each of which contains one or more readout
- This accounts for GPC1, and allows video guide cells with multiple readouts



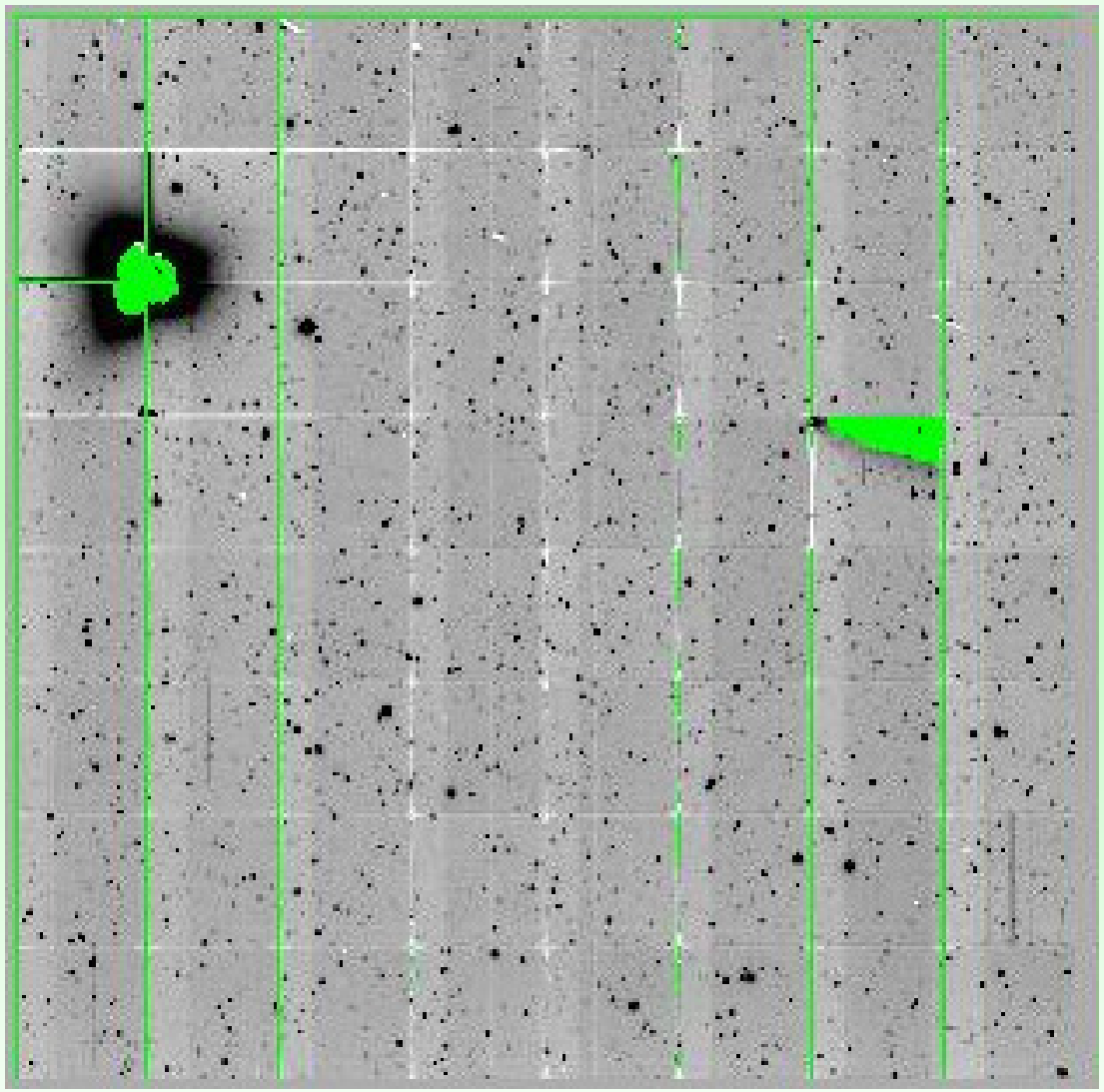
Make the code generic

- IPP code makes a minimum assumptions about the data as possible
- This general model allowed initial GPC2 observations to be reduced quickly
- From first light to MPC observatory code took four months
- Note that this image does not show the final devices for GPC2



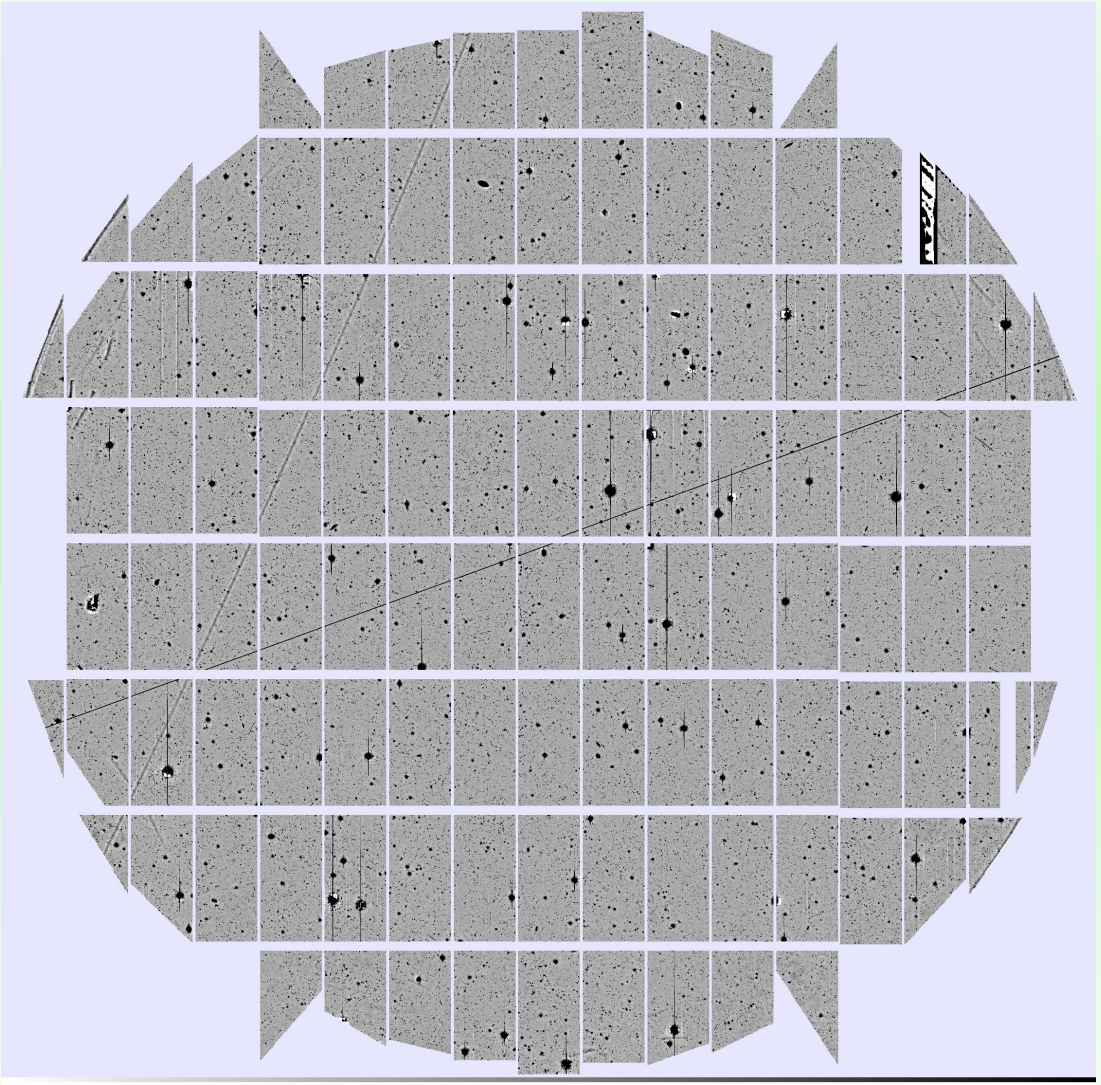
Make the code generic

- IPP code makes a minimum assumptions about the data as possible
- This general model allowed initial GPC2 observations to be reduced quickly
- From first light to MPC observatory code took four months
- Note that this image does not show the final devices for GPC2



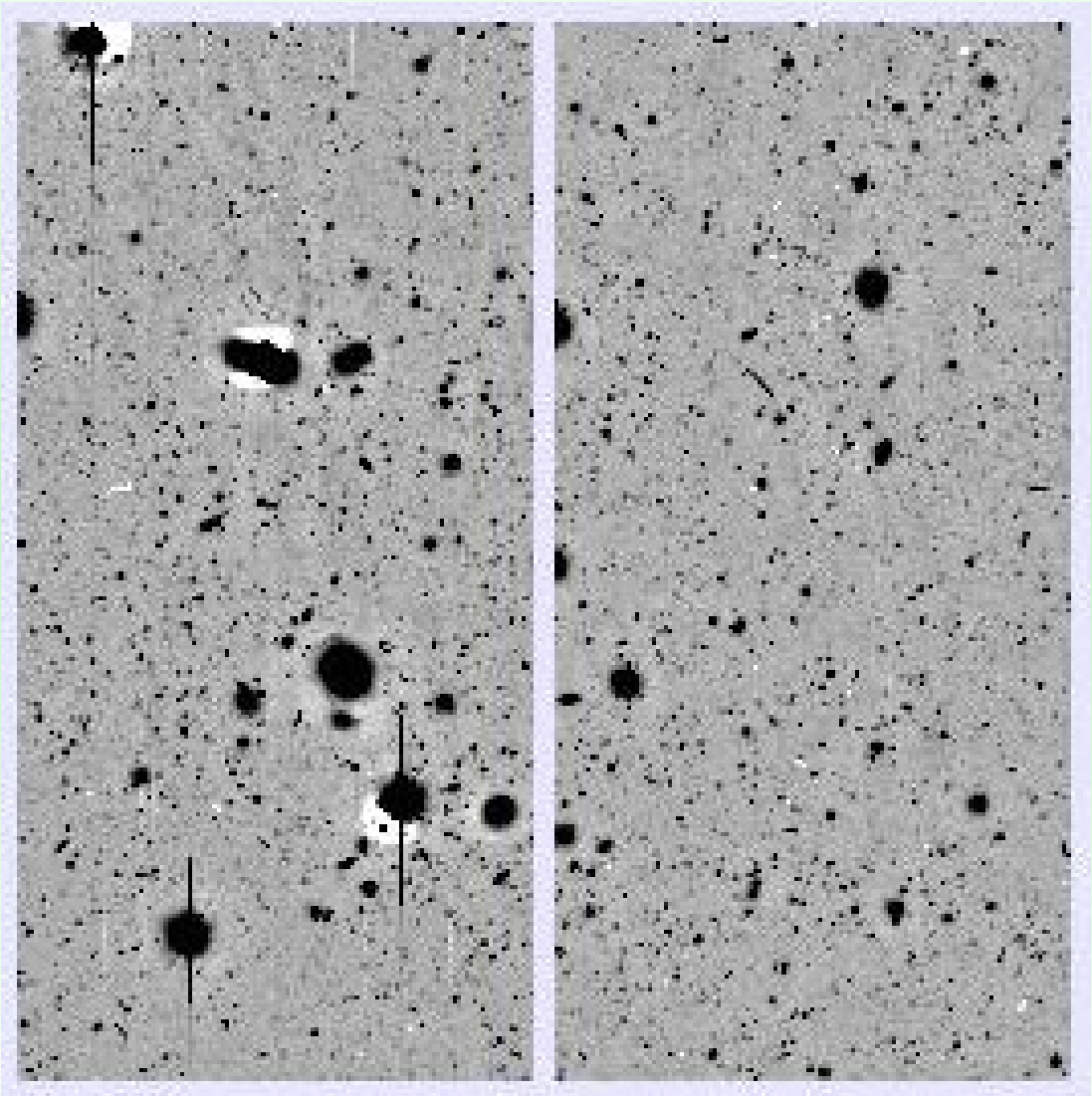
Make the code generic

- IPP code makes a minimum assumptions about the data as possible
- Subaru Hyper Suprime Cam matches this model as well
- Filters are sufficiently close to allow Pan-STARRS reference catalog to directly calibrate observations



Make the code generic

- IPP code makes a minimum assumptions about the data as possible
- Subaru Hyper Suprime Cam matches this model as well
- Filters are sufficiently close to allow Pan-STARRS reference catalog to directly calibrate observations



Make the processing generic

- Processing database stores information for each stage
- Modular pantasks scheduler queries database for potential jobs (load phase)
- Constructs command line for the job, assigns it to a processing host (run phase)
- Most jobs are perl script wrappers around C programs
  - C handles math/image analysis
  - perl handles configuration options, checking outputs, and populating database with results

Not limited to one location

- For PV3, we were offered access to the Mustang supercomputer at Los Alamos
- However, security reasons prevented direct access to the processing database
  - Moderate rewrite to bundle large numbers of jobs together, and "precalculate" the information needed
  - Pass simple list of C program calls to the MOAB scheduler
  - Return important files, database results
- Efficiency hit due to network transfers
- Roughly half of all chip-stack processing done remotely
- Duplicated for the University of Hawaii supercomputer for stack/forced warp photometry

Full Tests are Essential

- Component tests do not guarantee an operational pipeline
- Initial nightly processing saw bottlenecks due to competition between fast and slow jobs
- Reprocessing started with issues related to exposure inclusion
- Current data release is the internal “Processing Version 3”
- Previous versions resulted in initial science results, but also served as tests of the reduction system

Things Always Get Added

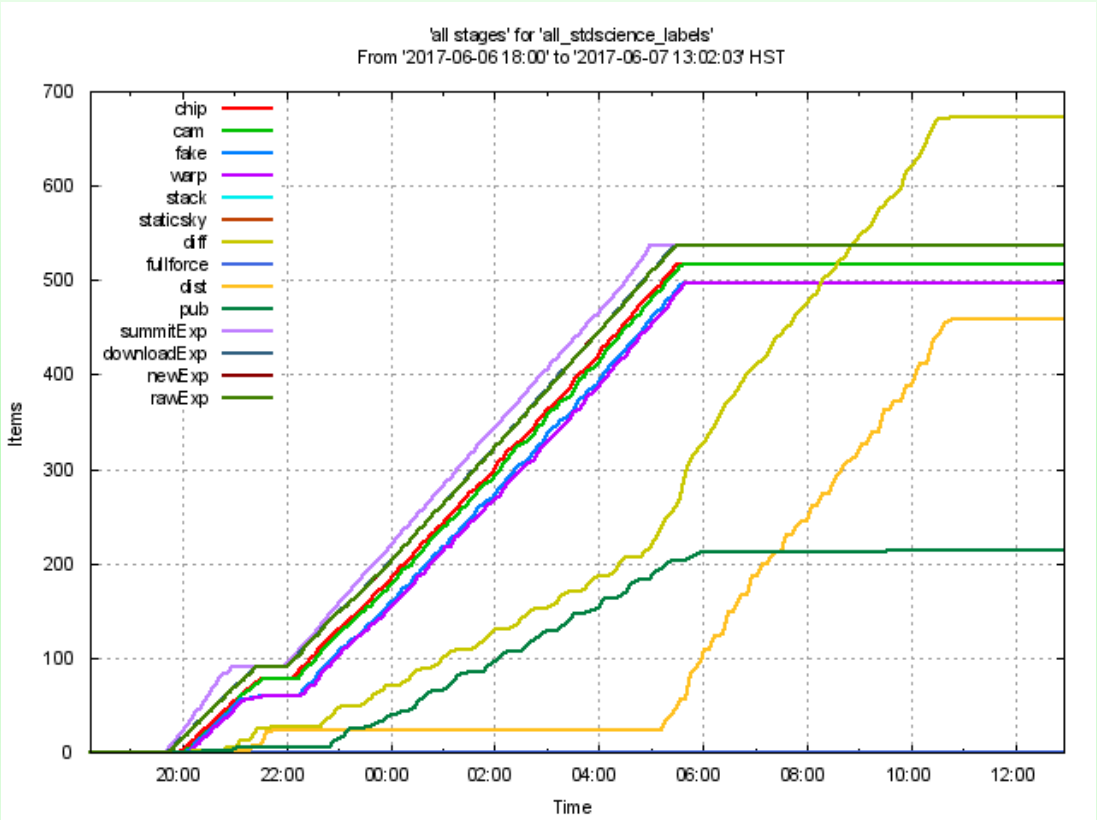
- Even still, additional stack photometry and forced warp was only included in PV3
- Previous iterations did not do complete warp-stack differences, and PV3 discovered issues with how the processing database handles these queries
- Adding those stages increases the time needed for the construction and calibration of the public database
- GAIA had their first data release, prompting another further recalibration

Oversight

- More eyes help find problems
- Important as the development team is small
- PS1 Science Consortium DRAVG meetings provided "external" view on the data, pointing out issues that needed to be resolved
  - Caught small (0.5 count) background bug
  - Prompted detailed false positive analysis
  - The skycal stage exists because stack zeropoints did not fully agree due to scatter in input seeing
  - Forced warp resolves the issue that stack PSFs are not uniform

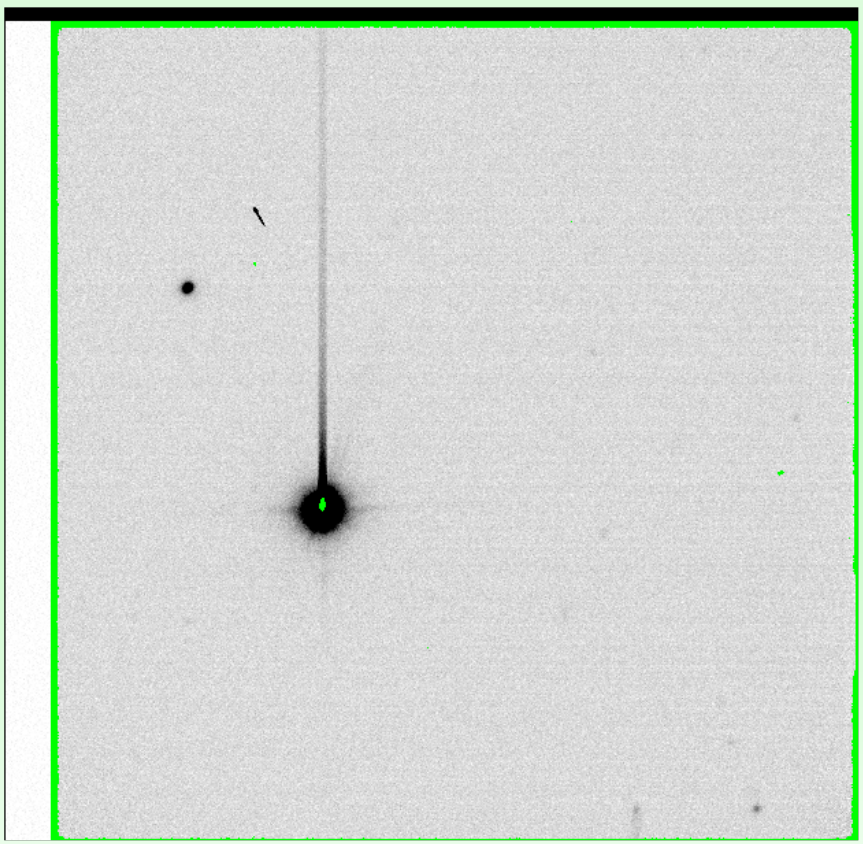
Monitoring

- Data volume is larger than can be easily inspected
- Nightly processing largely an issue of making sure that the lines all match
- Full sky reprocessing similar, ensuring that any patch of the sky currently being processed isn't older than others
- MD fields have issues ensuring that the optimum subset (seeing/zeropoint/etc) is included



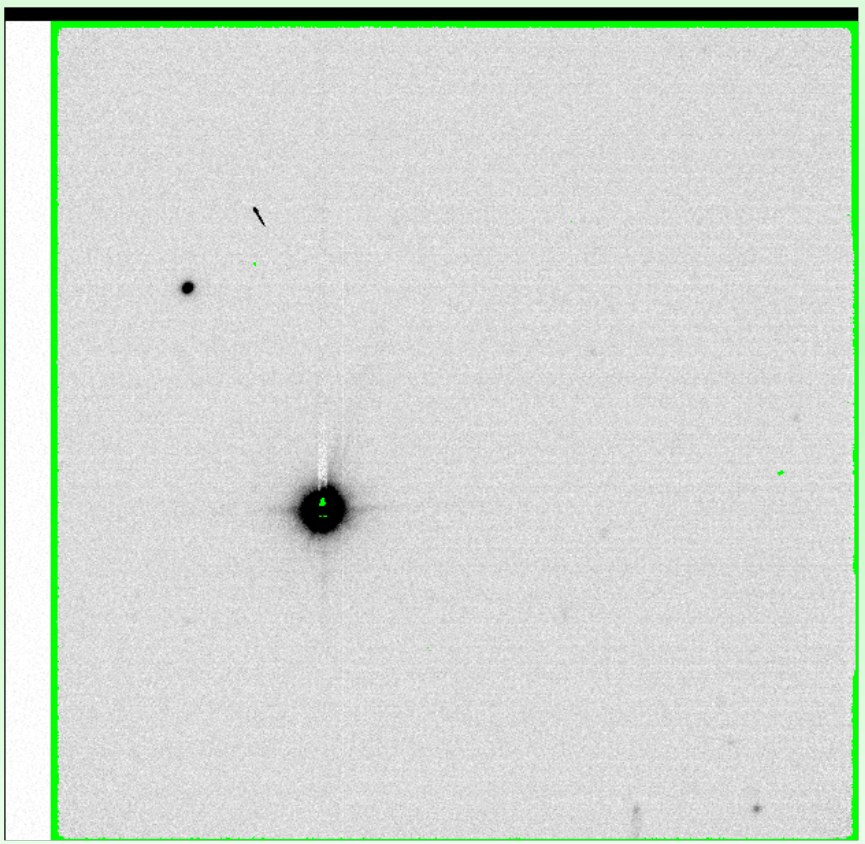
Nothing Works Perfectly

- Detector issues were problems early on
- Solutions developed to correct the most egregious
  - Burntool



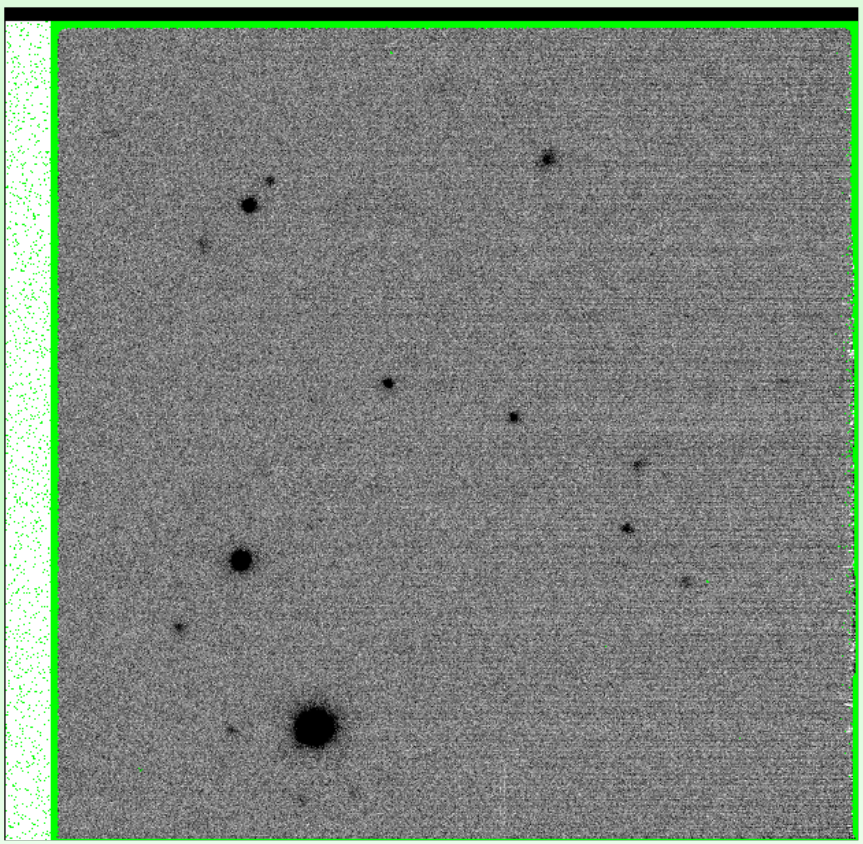
Nothing Works Perfectly

- Detector issues were problems early on
- Solutions developed to correct the most egregious
  - Burntool



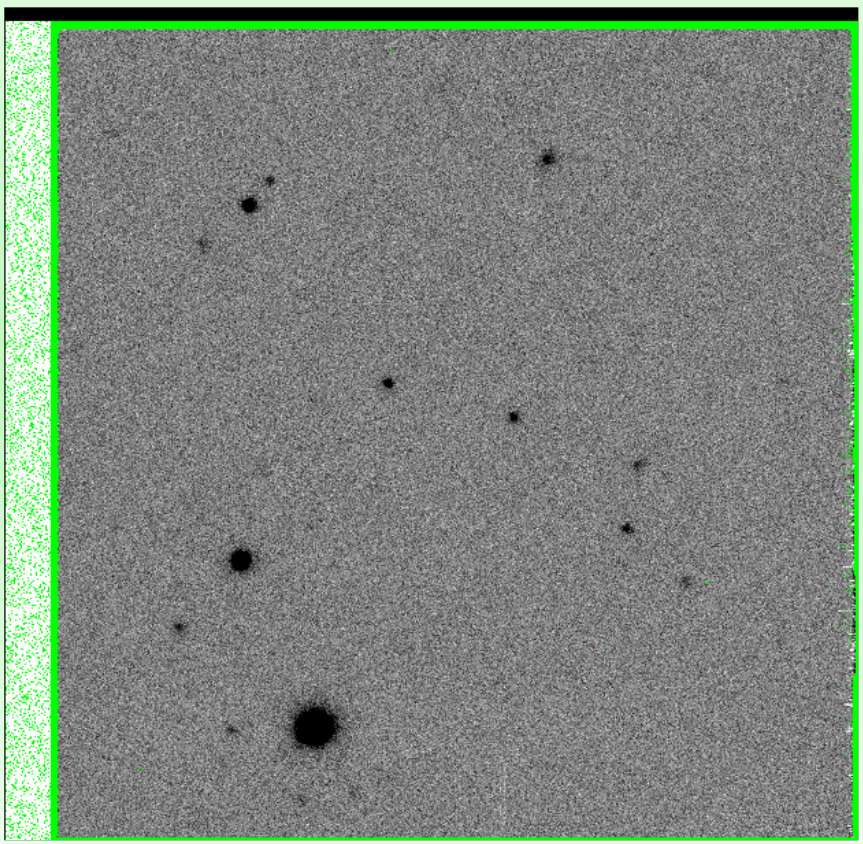
Nothing Works Perfectly

- Detector issues were problems early on
- Solutions developed to correct the most egregious
  - Burntool
  - Row-to-row bias



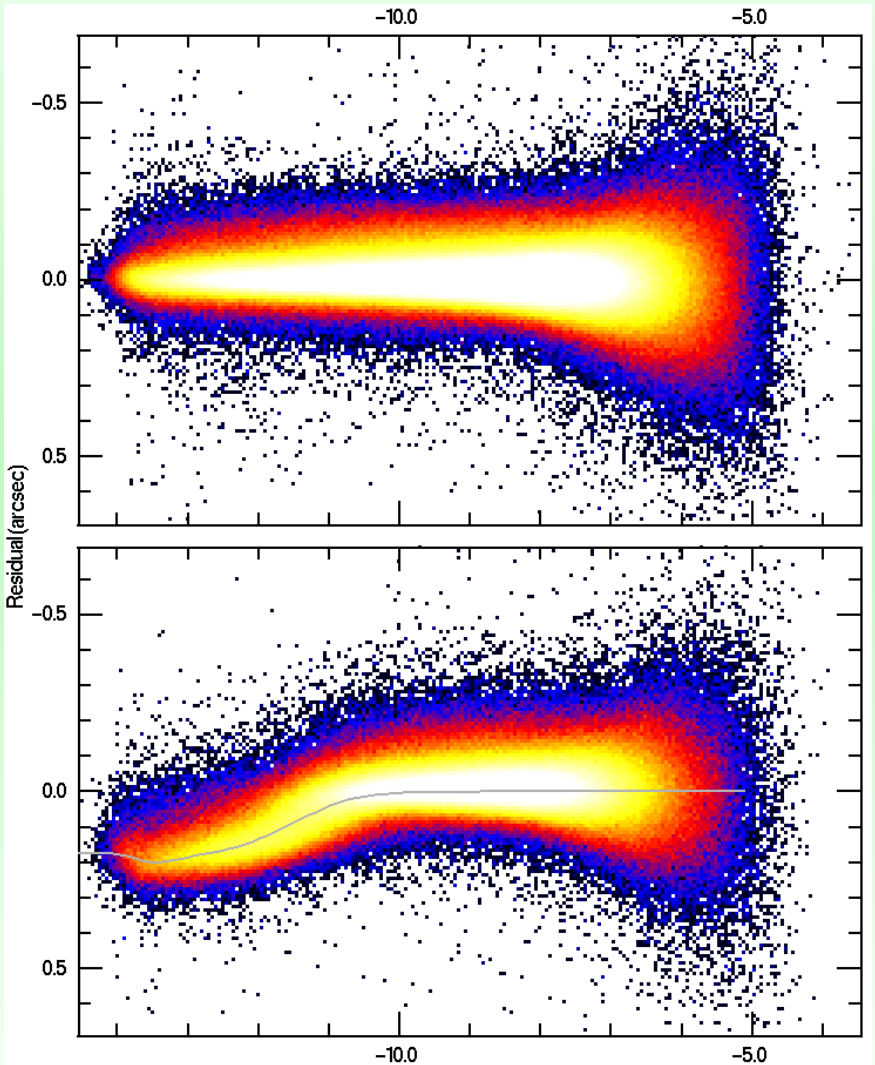
Nothing Works Perfectly

- Detector issues were problems early on
- Solutions developed to correct the most egregious
  - Burntool
  - Row-to-row bias



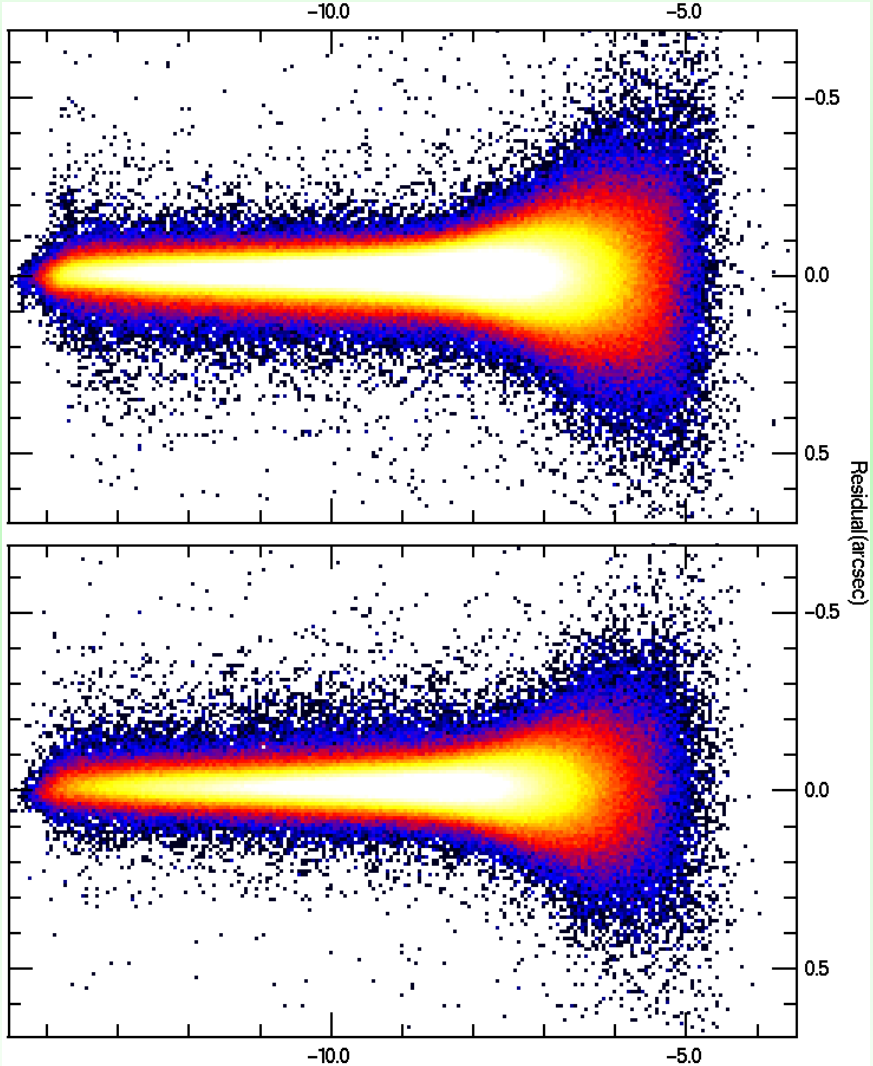
Nothing Works Perfectly

- Detector issues were problems early on
- Solutions developed to correct the most egregious
  - Burntool
  - Row-to-row bias
  - Koeppenhoffer effect



Nothing Works Perfectly

- Detector issues were problems early on
- Solutions developed to correct the most egregious
  - Burntool
  - Row-to-row bias
  - Koeppenhoffer effect
- Linear features triggered the satellite removal code that was required in USAF funding agreement
- This led to high mask fractions, which led to unhappy scientists



Moving the Cluster

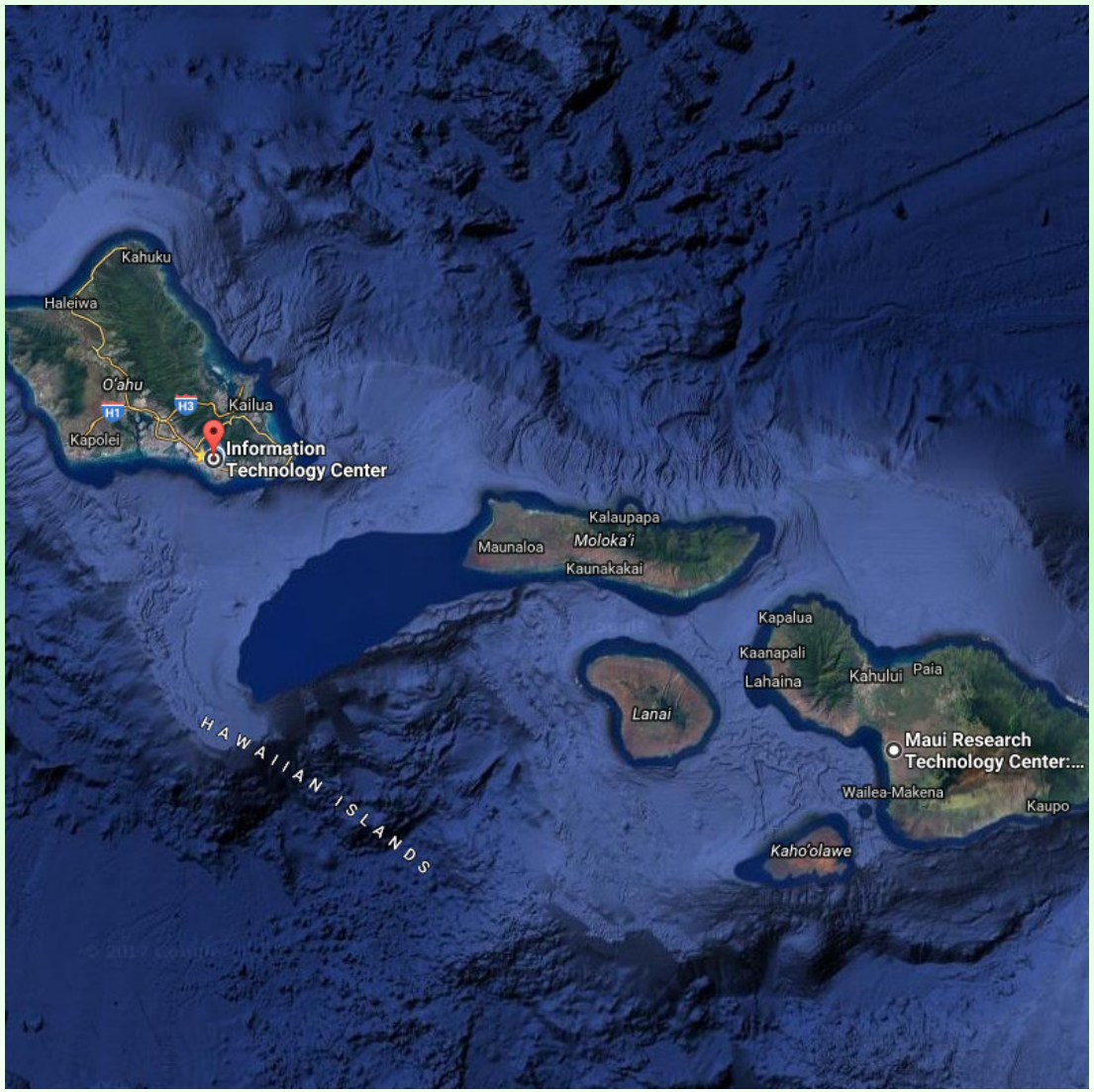
- Processing cluster initially hosted at the Maui High Performance Computing Center, operated by the Air Force
- After USAF funding ended (removing the masking restriction), moved to Maui Research Technology Center

Moving (Part of) the Cluster

- For the data release, we physically shipped twenty storage computers
- 2PB of image data, giving approximately 4GB/s transfer rate

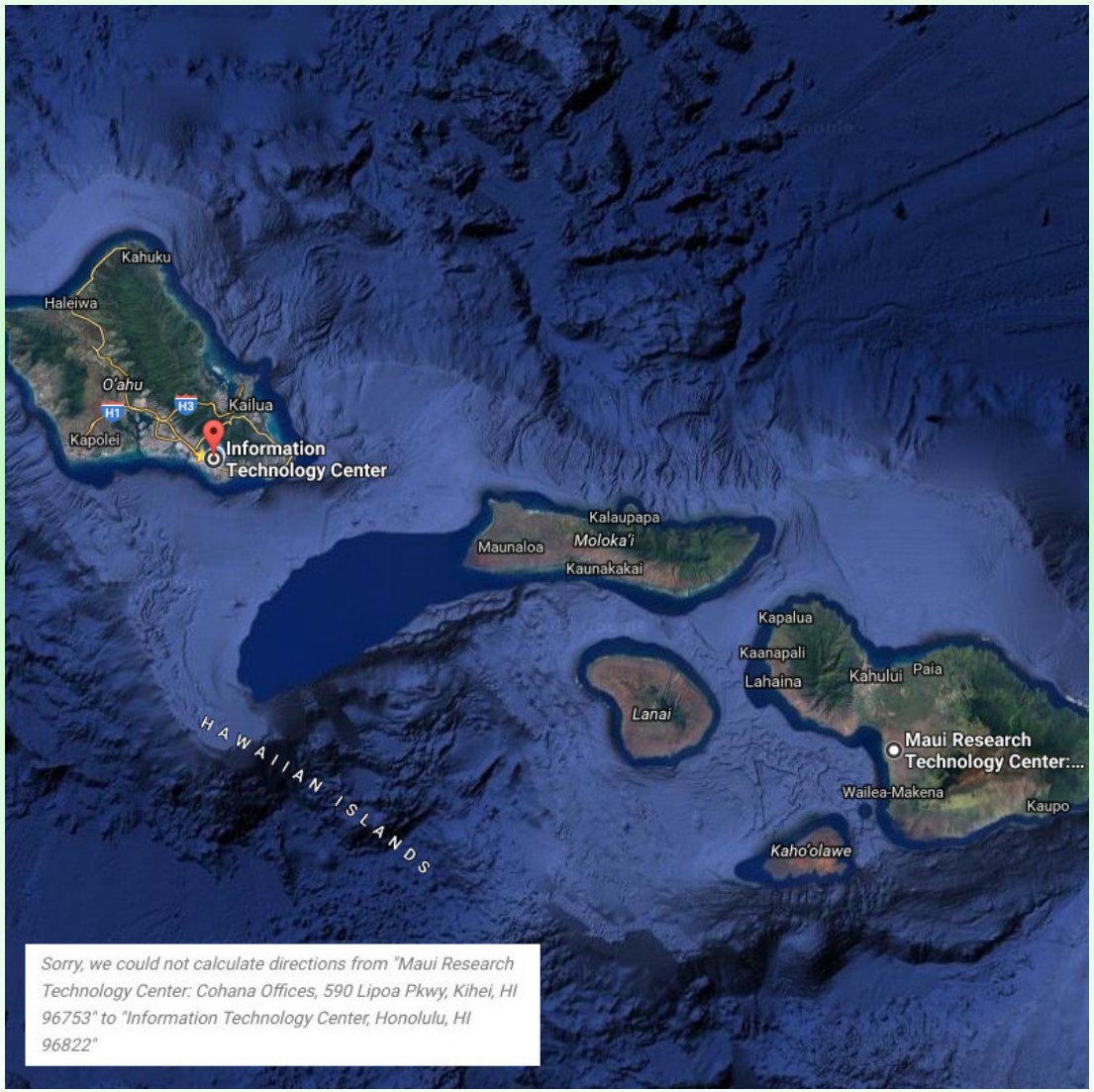
Moving the Cluster (Again)

- Recently moved entire cluster to new UH campus computing center
- Four stages of shipping to ensure fewest number of machines offline
- Able to maintain full nightly science processing during the move



Moving the Cluster (Again)

- Recently moved entire cluster to new UH campus computing center
- Four stages of shipping to ensure fewest number of machines offline
- Able to maintain full nightly science processing during the move



Disk space

- Data expands to fill the disk space available
- Have determined classes of data
  - Raw data: always have two copies, preferably geographically separated
  - Permanent data: catalogs and stacks that we expect to retain forever
  - Permanent metadata: PSF and background models, configuration status
  - Temporary data: image products that can be regenerated from the metadata
- Retiring old hosts requires ensuring that the essential data is preserved

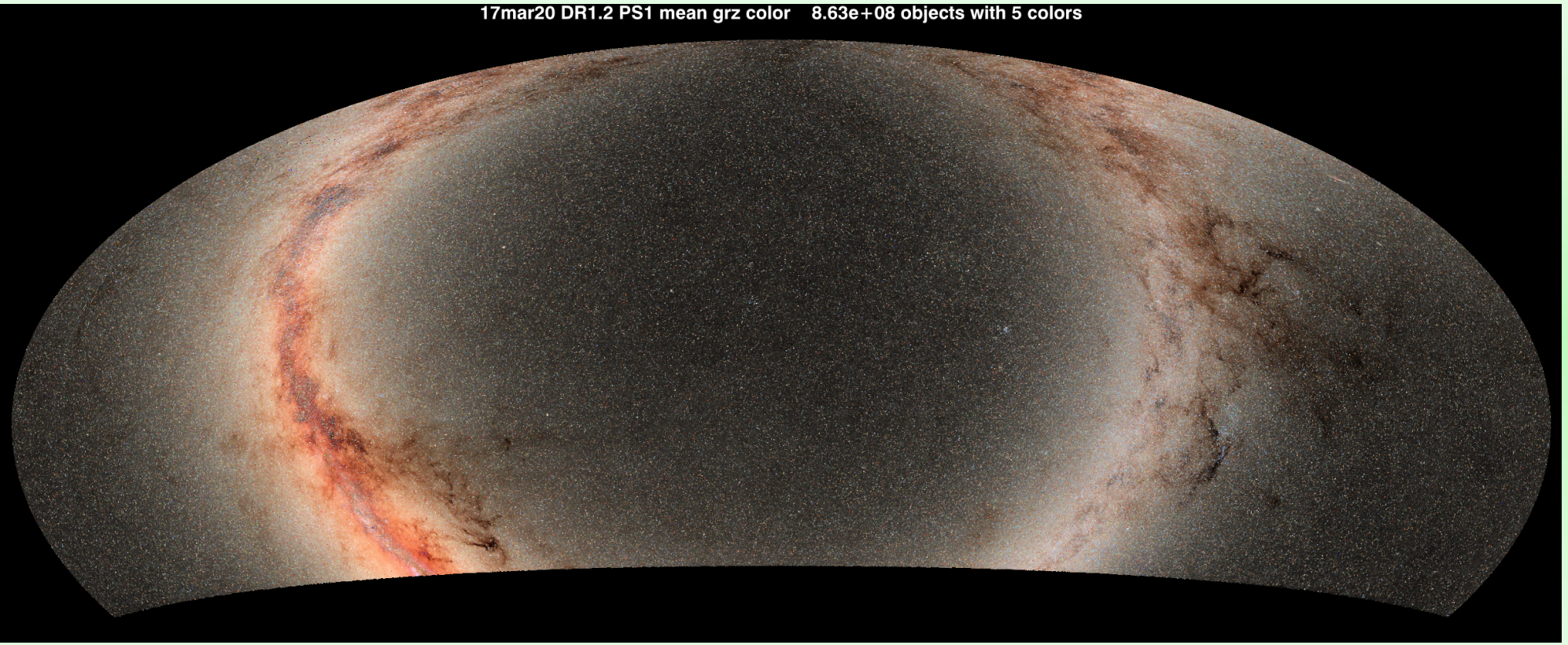
Documentation

- More documentation is always needed
- Public data release is the first step, as it gives a fixed target to describe
- Initial release papers are out, being updated and rewritten before submission
- STScI has helped turn those papers into documentation for the archive

The Future

- Data Release 2 will have updated calibration
  - All single epoch detections
  - All forced photometry measurements
  - All image warps
- PS2 with GPC2 will start a new round of commissioning, with a full focal plane of devices
- Nightly processing for the NEO search continues, and the detection efficiency will increase using both telescopes

Thank You


17mar20 DR1.2 PS1 mean grz color 8.63e+08 objects with 5 colors